LD score regression notes

Samuel Davenport

May 25, 2023

Let $Y = X\beta + \epsilon$, where $Y \in \mathbb{R}^n$ is a vector of phenotypes, $X \in \mathbb{R}^{n\times m}$ is the matrix of genotypes normalized to be mean zero and variance one 1, $\beta \in \mathbb{R}^m$ of effect sizes and $\epsilon \in \mathbb{R}^n$ is the error. In this model we assume that $\mathbb{E}(\epsilon) = 0$, $\text{var}(\epsilon) = (1-h^2)I$, $\mathbb{E}(\beta) = 0$ and $\text{var}(\beta) = (h^2/M)I$, where h^2 is the heritability. This is the same setting as in Bulik-Sullivan et al. (2015), see e.g. the beginning of their supplementary material.

For $1 \leq j \leq m$, let X_j be the jth column of X and let $\hat{\beta}_j = X_j^T Y/n$ and set $u_j = n\hat{\beta}_j^2$ be the χ^2 statistics. Moreover let

$$l_j = \sum_{k=1}^{m} r_{jk}^2 = \sum_{k=1}^{m} \mathbb{E}(X_{1j}X_{1k})$$

be the true LD scores. And let

$$\hat{l}_j = X_j^T X X^T X_j/n^2$$

be the estimates of the LD scores from the data. Since these may not be directly recorded let $\tilde{l}_j$ be an estimate of l_j from an independent reference dataset.

Then as I understand LD score regression fits the linear model (up to regression weightings),

$$u_j = a + \frac{n}{m}\tilde{l}_j h^2 + \eta,$$

where a represents the intercept term, and η the noise. This performing linear regression results in estimates $\hat{a}$ and $\hat{h}^2$ for the intercept and the heritability. Here importantly the estimates of the LD scores from the reference dataset are used instead of the actual values $(l_j)_{j=1}^m$ since these are unknown. Running this regression seems strange to me since it is based on the approximation

$$\mathbb{E}(u_j) \approx \frac{n}{m} l_j h^2 + Na + 1$$

that they derive in their paper. However because all of the u_js share the same X to me it doesn't seem possible to use them to infer on $\mathbb{E}(u_j)$ which is the expectation of u_j given that X can vary randomly. I.e. I would have thought you would need to have samples from the u_j distribution which had a different original X in order to be able to infer on $\mathbb{E}(u_j)$.

However I think that it would instead be possible to use the u_j to infer on $\mathbb{E}(u_j|X)$ since they share the same X. In particular the derivation in the supplementary of Bulik-Sullivan et al. (2015) implies that

$$\mathbb{E}(u_j|X) = n\text{var}(\hat{\beta}_j|X) = \frac{nh^2}{m}\hat{l}_j + 1 - h^2 = h^2\Big(\frac{n}{m}\hat{l}_j - 1\Big) + 1$$

Note that since the only dependence on X in this expression is via the $\hat{l}_j$ in fact $\mathbb{E}(u_j|X) = \mathbb{E}(u_j|\hat{l}_1, \ldots, \hat{l}_m) = \mathbb{E}(u_j|\hat{l}_j)$.

If the $\hat{l}_j$s were known it would thus make sense to instead run the regression

$$u_j = h^2\Big(\frac{n}{m}\hat{l}_j - 1\Big) + 1 + \eta$$

with a fixed intercept of 1 and noise error term η and solve to obtain an estimate of h^2 (also adjusting using regression weights to account for the dependence over j). I had thought (prior to your email) that, even though X was not known, the values $\hat{l}_j$ were stored. In fact that does not seem to be true which is a shame. Instead though I would propose to run the regression

$$u_j = h^2\Big(\frac{n}{m}\tilde{l}_j - 1\Big) + 1 + \eta$$

and solve for h^2. I would have thought a priori that this would do a better job than LD score regression because it tries to target $\mathbb{E}(u_j|X)$ rather than $\mathbb{E}(u_j)$. But if not I would like to understand what is better about LD score regression compared to this approach.

References

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.